Nagy Ágoston

Az automatikus szövegszegmentálás problémái

Bevezetés

A szövegek mondatokra (vagy legalábbis szövegszegmensekre) történő darabolása az NLP-rendszerek (Natural Language Processing, természetesnyelv-feldolgozás) egyik alapvető feladata. Több, természetes nyelvet kezelő alkalmazásnak végre kell hajtania ezt a szakaszt a tényleges szövegfeldolgozás és -manipuláció előtt. Ilyen alkalmazások például a szintaktikai elemzők (például INTEX-NOOJ), információkivonatolók, gépi fordítók, korpuszpárhuzamosítók. Vagy akár ide sorolhatnánk még a fordítást támogató szoftvereket, más néven a CAT-programokat (Computer-assisted Translation – géppel támogatott fordítás, tulajdonképpen fordítómemóriák): ide tartozik például a TRADOS és a Tranzit, melyek a kapott eredeti szöveget mondatokra bontják, és azokat (ember által véghezvitt) fordításaikkal együtt egy memóriába helyezik, így hasonló mondat esetén korábbi fordításainkból ötletet meríthetünk.[1]

Felmerülhet a kérdés, hogy ezek az alkalmazások miért a mondatot választják alapszegmensüknek. Több magyarázat is létezik: egyrészt a nagyobb egységeket nehezebb leírni, nehezebb számítógéppel kezelni. Másrészt a régi nyelvtani hagyományok miatt is, amelyek a mondatnak mindig is kiemelt szerepet nyilvánítottak a szövegen belül: a mondat szintjén alakul ki a predikáció (alany-állítmány összekapcsolódása), az ennél alacsonyabb egységekben (szó, szintagma) ez nyilván nem történik meg. Persze összetett mondatoknál ez a folyamat a tagmondat szintjén történik, de ahhoz is előbb a mondatot kell azonosítani.[2] (Fuchs 1993)

Mindebből az következik, hogy a szöveget gépi feldolgozása előtt célszerű mondatokra szegmentálni, azonban ez néha problémákba ütközhet. Jelen cikk célja többek között annak bemutatása, hogy a hagyományos „nagybetűvel

---

[1] Bár ezen alkalmazások hátrányára is válhat ez a túlzott mondatközpontúság. Ld. Farkas (2006).

[2] Természetesen bizonyos programoknak szükségük lehet mondathatárokon túl nyúló viszonyok felkutatására is. A gépi fordítóprogramoknak például hasznos lehet egy névmás antecedensének felderítése (ezt a szakirodalom *anaphora resolution* névvel illeti) a megfelelő fordításhoz (Mitkov 2003). Tekintsük az alábbi példamondatot:

(i) *Regarde cette chemise$_i$! Qu'elle$_i$ est belle!*

Rögtön észrevehetjük, hogy a mondat szó szerinti angol fordítása nem a megfelelő névmási alakkal térne vissza, hiszen az *elle* névmásnak az angolban a *she* felel meg. Érdekes módon azonban a Google fordítóporgramja helyesen adta vissza ennek a mondatpárnak a fordítását:

(ii) *Look at this shirt! How it is beautiful!*

kezdődik és ponttal, felkiáltójellel, kérdőjellel végződik"-elv igencsak pontosításra szorul.

A szövegszegmentálás problematikáinak bemutatásához egy saját programkód futásai során létrejött eredményeket is felhasználunk. A példák egy részét Stendhal *Vörös és fekete* című regényének magyar (továbbiakban VF) és francia nyelvű eredeti változata (*Le rouge et le noir* – továbbiakban RN) adta. Azért ez a két szöveg, mert mindkettő letölthető egyszerű szöveges formátumban (*.txt), így természetesen könnyebb őket feldolgozni. A szövegek szegmentálását a Perl nevű programmal végeztük, mely a leginkább alkalmas programnyelv szövegek manipulálására. Kezelése sem túl bonyolult, mondhatni a C nyelv egyszerűsített változata, de a Perl Windows alatt is működik (bár csak az MS-DOS parancssorból), míg a C leginkább csak Linux alatt futtatható.

Az első részben a mondat fogalmának tradicionális megközelítéseit vizsgáljuk, azaz azt, hogy a nyelvtanok általában hogyan viszonyulnak a mondat fogalmához. Az ezután következő részekben arra fektetünk nagy hangsúlyt, hogy milyen problémákba ütközhet az automatikus szövegszegmentálás, ezen belül is a mondathatárt nem mindig jelölő írásjelekkel foglalkozunk, végül azt mutatjuk be, hogy jelenleg milyen automatikus szövegszegmentálási tendenciák léteznek ezen problémák kiküszöbölésére.

1. A mondat fogalmának tradicionális megközelítései

A hagyományos, akár előíró, akár leíró nyelvtanok (például Riegel et al. 1994, Arrivé et al. 1986), általában három olyan kritériumot említenek, melyek segítségével a mondatok egy adott szövegen belül behatárolhatók. Az első a grafikai kitétel, mely szerint minden mondat egy olyan szósorozat, melyet bal oldalról kezdő nagybetű, másik oldalról pedig mondatzáró központozás zár. Hangtani szempontból a mondat olyan hangok sorozata, melyeket két szünet határol (egy az elején és egy a végén), és az adott mondat típusának megfelelő intonáció jellemez (kérdő, állító vagy felszólító). A harmadik kitétel szerint, mely inkább a jelentésre fekteti a hangsúlyt, a mondat egy „gondolati egységet" alkot, következésképp nem minden szósorozat mondat.

A felhozható ellenpéldák száma szinte végtelen, ezt ki is használva, a nyelvtanok írói rengeteg ilyen esetet említenek a mondat problematikájának felvetésekor. Ugyanakkor mi magunk is könnyen beláthatjuk, hogy ezek a kritériumok nem elegendőek. Először is, annak bizonyításául, hogy a központozás nem elegendő a mondat határainak meghatározásához, rengeteg olyan mondatot találhatunk, melyek ugyan tartalmaznak felkiáltójelet vagy kérdőjelet, de ez utóbbiak mégsem az adott mondat határait jelölik. Ezt az alábbi példa is jól szemlélteti:

(1) *a)* Quelle différence! reprit vivement le geôlier. (RN, 3.fejezet)
b) Ó, az nagy különbség! – vágott közbe a börtönőr. (VF, 3.fejezet)

Tulajdonképpen hány mondatról is beszélhetünk ebben a szósorozatban? Ezen kívül, a beszéd során (mondathatárt nem jelölő) szünetek az adott mondaton belül többször is előfordulhatnak, mint például a francia diszlokált szerkezetekben:

(2) *Pierre, le chocolat, il l'aime.*

És végül mit értsünk „gondolati egységen"? A „szemantikai teljesség" alapján nem lehet mindig eldönteni, hogy az adott szósorozat egy vagy több mondatot alkot. Ha a „Hideg van. Péter nem fog jönni." két mondat, tehát két gondolati egység, akkor a „Hideg van, Péter tehát nem is fog jönni." miért csak egy?

Minthogy most írott korpuszokkal foglalkozunk, a fonetikai kitételt semmiképpen sem alkalmazhatjuk szövegek szegmentálásra. A szemantikai definíció sem alkalmazható igazából esetünkben, hiszen mint ahogy az előbbiekben láthattuk, a „gondolati egység" fogalma még az átlagos nyelvhasználók számára sem mindig írható pontosan körbe, hát még egy számítógép esetében, mely számára az értelmezés még nehezebb feladat. Azt ki is jelenthetjük, hogy az egyszerű számítógépes nyelvészeti alkalmazások (például korpuszpárhuzamosítók) értelmezéssel nem foglalkoznak. Az értelmezéshez, és az egyéb, emberi intelligenciához kapcsolódó képességek (döntéshozás, érvelés, vagy a pragmatikából jól ismert implikációk értelmezése) a mesterséges intelligencia hatáskörébe tartoznak (Fuchs 1993). Másrészről a szemantikai kritérium számítógépes implementációja erősen nyelvfüggő, és valószínűleg sokkal időigényesebb is (mármint informatikai viszonylatban). Így csak a tipográfiai kritérium maradt, mely annyiból szerencsés, hogy implementálása viszonylag könnyebb, azonban a szegmentált korpuszt utána gyakran célszerű leellenőrizni.

2. Szövegszegmentálási kísérletek a tipográfiai kritérium alapján

Sokan gondolják, hogy a mondatszegmentálás egy viszonylag egyszerű feladat, elég csak a nagybetűket figyelni, valamint a központozást. Sajnos a valóságban ez nem mindig igaz. A hagyományos nézőpont implementációja is lehet sikeres, de ez eléggé korpuszfüggő: a Brown-korpuszon a hibaaránya például 6,5%-os volt, ami nem jelentős (persze nézőpont kérdése), de ha lehet, miért ne pontosítanánk rajta? (Mikheev 2003)

Vizsgáljuk meg, milyen problémák merültek fel a két szövegünk esetében: Stendhal világszerte ismert regényének francia eredeti verziójában, és annak magyar nyelvű fordításában. Ezen kívül a továbbiakban említést teszünk még olyan jelenségekre, melyek ugyan az adott szövegekben nem szerepeltek, azonban mégis bármikor előfordulhatnak.

Gyakran előfordult – mind a magyar, mind a francia változatban –, hogy egy adott, általában mondatzáró, központozási jel mégsem a mondat végét jelölte. Erre már az előző alfejezetben is említést tettünk. Felmerülhet a kérdés, hogy az alábbi mondatokat érdemes-e felbontani vagy sem:

(3) *a)* Tu approuves donc mon projet? dit M. de Rênal, remerciant sa femme, par un sourire, de l'excellente idée qu'elle venait d'avoir... (RN, 3. fejezet)
b) Tehát helyesled a tervemet? – kérdezte de Rênal úr, s mosolyogva nyugtázta az asszony kitűnő megjegyzését... (VF, 3. fejezet)

(4) *a)* – N'avez-vous point d'autre nom que Julien? lui dit-elle. (RN, 14. fejezet)
b) – Julienen kívül nincs más neve? – kérdezte tőle. (VF, 14. fejezet)

Számunkra természetes, hogy a közbeékelődő mondatokat nem választjuk el, azonban miután a program a szabályos kifejezésünk első verzióját ráillesztette az első tagmondatokra, ott jogosan szét is bontja a mondatot. Célszerű volt tehát úgy módosítani a programkódot, hogy egészen addig tart a mondat, amíg központozást záró jel után szóköz és kisbetű, vagy szóköz és kötőjel (legalábbis a magyarban), stb. következik (egyszerűbben megfogalmazva: a nagybetűt kivéve bármi követhet még mondathatárt jelölő írásjelet, és az ehhez a karakterhez tartozó szósorozat szintén olyan központozással végződik, mely után nem áll nagybetű, és így tovább).

Különösen a magyar verzióban volt gyakori, hogy célszerűvé vált a kettőspontot is felvenni a mondatzáró írásjelek közé, de találhatunk erre francia példát is:

(5) *a)* Si, en entrant à Verrières, le voyageur demande à qui appartient cette belle fabrique de clous qui assourdit les gens qui montent la grande rue, on lui répond avec un accent traînard: *Eh! elle est à M. le maire.* (RN, 1. fejezet)
b) És ha Verrières-be érve megkérdi, kié ez a szép szögsajtoló, amely megsüketíti a főutcai járókelőket, a verrières-iek vontatottan felelik:
– Hát a polgármester úré... (VF, 1. fejezet)

Azonban időnként a kettőspontot követő nagybetű nem biztos, hogy mindig újabb mondatot jelöl. A következő példa azt szemlélteti, hogy a kettőspont után álló tulajdonnév semmiképpen sem alkothat külön mondatot – már csak a zárójel miatt sem (azonban az egyik fordítást támogató szoftver használatakor már találkoztam hasonlóval, így ez még talán nem is okoz komolyabb gondot, főleg mivel csak egyszer fordult elő).

(6) Így az a fűrészmalom is, amelyik [...] a Verrières-be érkező utasnak rögtön szemébe tűnik (a tető fölött magasba helyezett deszkán óriási betűk hirdetik a tulajdonos nevét:

Sorel), – hat évvel ezelőtt még ott állt, ahol [...] (VF, 1.fejezet)

A szövegszegmentáló rendszerek számára a legnagyobb problémát mindig is a rövidítések jelentették, melyek azért különösen veszélyesek, mert a rövidítés végén (vagy akár közepén többször is) előforduló pont semmiképpen sem jelezheti a mondat végét. A regény francia változatában egyetlen ilyen rövidítés szerepelt: a *M.* (*Monsieur*). Ez a rövidítés azért is problémás, mert utána gyakran szerepel olyan nagybetű, mely nem a mondat határát jelzi, például *M. Ducrest* (Ducrest Úr). Erre a legegyszerűbb megoldást mindenképpen az jelentené, hogy az *M* karaktert kizárjuk a mondatvégi pont előtt. Azonban ez eléggé nyelvfüggő, hiszen a magyar nyelvre lefuttatva a programkód az alábbit is eredményezte:

(7) <3320> A XIX. </3320>
<3321> században viszont a férj a közmegvetés csapásával öli meg feleségét; az asszony előtt azonnal bezárulnak a szalonok. </3321> (VF, 21.fejezet)

Itt a <szám> az adott számú mondat elejét, míg a </szám> az adott mondat végét jelöli. Látható, hogy a magyar nyelv esetében a római számokat is kell iktatni a mondatvégi pont előtt. Sőt, ha jobban belegondolunk, még az arab számokat is, hiszen dátumok esetében sem illő szétválasztani például az *1986. december 26*-át sem.

De vajon megállhatunk-e itt? Egy egyszerű irodalmi szöveg esetén biztosan, de nézzük mi történhet egy átlagos, interneten is olvasható szöveg esetén:

(8) Un O.V.N.I. est apparu dans le ciel de Marseilles.

A pontok miatt ezt egy szövegszegmentáló akár négy mondatra is bonthatná, amiből már látható, hogy az eddig megszorítások semmiképpen sem elegendőek. Első megoldásként bevezethetjük, hogy nagybetű nem szerepelhet mondatvégi írásjel előtt; ezzel ugyan a problémák nagy részét kiszűrtük, de ha például egy szöveget csupa nagybetűsként kap bemenetként egy fordítást támogató program (például TRADOS), akkor azt egy szegmensként kezelje a program?

Másik, már kissé jobban működő eljárás lehet, ha azt mondjuk, hogy a mondatvégi írásjelet mindig vagy szóköz vagy sorvége jel kövessen (a C-ben ez '\n', Perl-ben ez utóbbi vagy $), minthogy a több iniciálét tartalmazó rövidítések pontjai után, megszokott hagyomány szerint, sosem teszünk szóközt (például S.N.C.F.), a mondatvégi pont és egyéb írásjelek (pl. ; , stb.) után pedig mindig. Azonban ez a módszer egy-egy szóközkihagyás vagy -betoldás esetén eléggé félrevezető eredményeket produkálhat (bár ennek valószínűsége kicsi).

A szövegszegmentáló alprogram megfelelő működéséhez a legtöbb alkalmazás egy rövidítésszótárt használ, melynek segítségével a program ellenőrizni tudja, hogy az adott pont rövidítésben szerepel, vagy mondatot határol (Mikheev 2003). Azonban aki a francia kultúrát akár csak alapszinten ismeri, tudja, hogy

– főleg a latin nyelv maradványaként – se szeri se száma a rövidítéseknek, amelyeket ugyan fel lehet venni egy szótárba, de a szótár félévenkénti frissítése sem lehet elegendő a tömegével felbukkanó újabb és újabb francia mozaikszavak tárolásához. Elég csak a tavaly nagy vihart kavart C.P.E.-re gondolni (Contrat Première Embauche), mely lényegében a pályakezdő fiatalok munkahelyi próbaidejét igyekezett két hónapról két évre kitolni: ez a mozaikszó például a megjelenése utáni hetekben szabályosan elárasztotta a sajtót.

Ezeken a jelenségeken kívül még rengeteg olyan esetről számolhatnánk be, amikor a mondat behatárolása pusztán tipográfiai kritériumok alapján nem elegendő. Ezek közé tartozik például az alábbi szöveg is:

(9) Washington s'inquiète et réclame l'éviction de trois officiers – proches du président – impliqués dans des atteintes aux droits de l'homme en... 1991. (Mourad 1999: 9)

A három pont után itt egy szóköz és szám következik, mégsem beszélhetünk két mondatról, mert az évszám a mondat szerves részét képzi.

3. Kitekintés

Mint ahogy az eddigiek során is láthattuk, a mondatokra történő (precíz) szegmentálás korántsem olyan egyszerű, mint ahogy az első látásra tűnik. Éppen ez okból kifolyólag több módszert is kifejlesztettek már e cél megvalósítására. A legtöbben egyetértenek azzal, hogy a hagyományos megközelítés önmagában nem elegendő, és mindenképpen célszerű egyéb szótárakkal illetve kontextuselemzőkkel az adott programot kiegészíteni.

A problémára a szakirodalmak legtöbbször az SBD (Sentence Boundary Disambiguation – Mondathatár-egyértelműsítés) kifejezéssel utalnak. Alapvetően kétféle SBD-alkalmazás létezik: az egyik típusba tartozó alkalmazások szabály alapúak, míg a másik típusba tartozóak statisztikai módszerekkel dolgoznak. A szabály alapú rendszerek kézzel, ember által megírt szabályokkal (ún. szabályos kifejezésekkel) dolgoznak, melyeket rövidítés-, ill. tulajdonnévszótárral egészítenek ki. Az általunk kidolgozott rendszer is szabályos kifejezések alapján működik: egy ilyen rendszert könnyű létrehozni, azonban ezt tökéletesíteni már sokkal nehezebb és időigényesebb. Ezen kívül, ezek a rendszerek erősen nyelv- és korpuszfüggőek is, tehát nem minden területen lehetnek kompatibilisek – általában csak azon a területen, amelyre azt kidolgozták. Hibaarányuk sem csekély, általában 4% körüli értékekről beszélhetünk. (Mikheev 2003)

Ilyen típusú szövegszegmentáló programot hozott létre például Anne Dister, a liège-i egyetem munkatársa, aki az INTEX-rendszer segítségével, transzducerek (olyan automaták, melyek beolvassák a szöveg egy szakaszát, és az ehhez a szöveghez társított információ segítségével döntik el, hogy az adott „.!" vagy „?"

mondatvégét jelöl-e) használatával kísérelte meg a szövegek szegmentálását. Ide sorolható még a szintén közismert SATZ is (németül mondatot jelent), mely megfigyelte még a mondatzáró központozási jelöltek bal és jobb oldali kontextusát is, tehát lexikális modult is épített be rendszerébe. (Mourad 1999)

A statisztikai rendszerek alapját azonban általában az képzi, hogy az alkalmazás saját maga tanulja meg a szabályokat általában előre szegmentált korpuszok alapján; mindamellett ma már olyan rendszerek is megjelentek, melyek előre nem szegmentált, nyers szövegekből is képesek szabályokat kinyerni. Ezek a rendszerek azt használják ki, hogy a központozás csak nagyon kevés esetben ad kételyre okot, így rengeteg szabályt kinyerhetünk az egyértelmű központozás alapján. Ilyen rendszerek közé tartozik például az LTSTOP. Ezek a rendszerek már kevésbé nyelvfüggőek, bár nyelvi bővítményekkel hatékonyságuk igencsak megnövelhető. Ezen alkalmazások hibaaránya átlagosan 0,8–1,5% között van – tehát jóval kevesebb, mint a szabályalapú modelleké. (Mikheev 2003)

Összegzés

Jelen tanulmány célja annak bemutatása volt, hogy a mondatokra (vagy legalábbis szövegszegmensekre) történő darabolás megvalósítása korántsem olyan egyszerű feladat, mint amennyire az első látásra tűnik. Márpedig ez a lépés fontos minden számítógépes nyelvészettel kapcsolatos alkalmazás (például fordítást támogató programok, szintaktikai elemzők és annotálók stb.) számára, hiszen a mondatokra történő szegmentálás után kezdődhet csak el igazán az adott szöveg manipulálása (például automatikus fordítása).

Mint ahogy láthattuk, a hagyományos nézőpont – mely szerint minden olyan szósorozat, mely nagybetűvel kezdődik és ponttal, kérdőjellel, vagy felkiáltójellel végződik, az mondat – nem mindig helytálló, némi pontosításra szorul. Azonban még az ebben az irányban továbbfejlesztett szabály alapú alkalmazások is hordozhatnak magukban hibalehetőséget. Ezért a jelenlegi tendencia nem a szabályalapú, hanem a statisztikai (vagy öntanuló) mondatszegmentálók folyamatos terjedése. Ez utóbbiak esetében a hibaarány is jelentősen kisebb, valamint a legújabb rendszerekben már nincs is szükség előre annotált korpuszokra sem.

Bibliográfia

Arrivé, M., Gadet, F., Galmiche, M. 1986. *La grammaire d'aujourd'hui: Guide alphabétique de linguistique française.* Paris: Flammarion.

Farkas, I. 2006. L'eurhongrois: du jargon indoeuropéen? In: Gyimesi, T., Koncz, B., Kőrös, A., Kovács K., Pálfy, M. (szerk.) *«prismes irisés»*. Szeged: Klebelsberg Kunó Egyetemi Kiadó. 409–414.

Fuchs, C. 1993. *Linguistique et Traitement automatique des langues.* Paris: Hachette.

Mikheev, A. 2003. Text Segmentation. In: Mitkov, R. szerk. *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press. 201–218.

Mitkov, R. 2003. Anaphora Resolution. In: Mitkov, R. szerk. *The Oxford Handbook of Computational Linguistics.* Oxford: Oxford University Press. 266–283.

Mourad, G. 1999. *La segmentation de textes par l'étude de la ponctuation*; Actes de colloque international, CIDE'99. Damas, Syrie. 155–171.

Riegel, M., Pellat, J-C., Rioul, R. 1994. *Grammaire méthodique du français.* Paris: PUF.

Források

Stendhal. *Vörös és fekete.* (Ford: Illés Endre)
http://www.mek.iif.hu/porta/szint/human/szepirod/kulfoldi/stendhal/vorosf/
Stendhal. *Le rouge et le noir.* 1830.
http://abu.cnam.fr/cgi-bin/donner?rouge1